SinGAN-Seg: Synthetic training data generation for medical image segmentation

Year: 2021 | Citations: 103 | Authors: Vajira Lasantha Thambawita, Pegah Salehi, Sajad Amouei Sheshkal, Steven Hicks, Hugo

Abstract

Analyzing medical data to find abnormalities is a time-consuming and costly task, particularly for rare abnormalities, requiring tremendous efforts from medical experts. Therefore, artificial intelligence has become a popular tool for the automatic processing of medical data, acting as a supportive tool for doctors. However, the machine learning models used to build these tools are highly dependent on the data used to train them. Large amounts of data can be difficult to obtain in medicine due to privacy reasons, expensive and time-consuming annotations, and a general lack of data samples for infrequent lesions. In this study, we present a novel synthetic data generation pipeline, called SinGAN-Seg, to produce synthetic medical images with corresponding masks using a single training image. Our method is different from the traditional generative adversarial networks (GANs) because our model needs only a single image and the corresponding ground truth to train. We also show that the synthetic data generation pipeline can be used to produce alternative artificial segmentation datasets with corresponding ground truth masks when real datasets are not allowed to share. The pipeline is evaluated using qualitative and quantitative comparisons between real data and synthetic data to show that the style transfer technique used in our pipeline significantly improves the quality of the generated data and our method is better than other state-of-the-art GANs to prepare synthetic images when the size of training datasets are limited. By training UNet++ using both real data and the synthetic data generated from the SinGAN-Seg pipeline, we show that the models trained on synthetic data have very close performances to those trained on real data when both datasets have a considerable amount of training data. In contrast, we show that synthetic data generated from the SinGAN-Seg pipeline improves the performance of segmentation models when training datasets do not have a considerable amount of data. All experiments were performed using an open dataset and the code is publicly available on GitHub.